Molecular population genomics: a short history

Genet. Res., Camb. (2010), **92**, *pp.* 397–411.
doi:10.1017/S0016672310000522

Molecular population genomics: a short history

BRIAN CHARLESWORTH*
Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK

(*Received 28 September 2010 and in revised form 22 October 2010*)

Summary

Population genomics is the study of the amount and causes of genome-wide variability in natural populations, a topic that has been under discussion since Darwin. This paper first briefly reviews the early development of molecular approaches to the subject: the pioneering unbiased surveys of genetic variability at multiple loci by means of gel electrophoresis and restriction enzyme mapping. The results of surveys of levels of genome-wide variability using DNA resequencing studies are then discussed. Studies of the extent to which variability for different classes of variants (non-synonymous, synonymous and non-coding) are affected by natural selection, or other directional forces such as biased gene conversion, are also described. Finally, the effects of deleterious mutations on population fitness and the possible role of Hill–Robertson interference in shaping patterns of sequence variability are discussed.

1. Introduction

Population genomics is a new term for a field of study that is as old as the field of genetics itself, assuming that it means the study of the amount and causes of genome-wide variability in natural populations. Indeed, the problem of characterizing natural variability greatly concerned Charles Darwin, since evolutionary change under natural selection requires the existence of heritable variation in the traits in question. By drawing on evidence from domesticated species of animals and plants, Darwin succeeded in demonstrating the existence of such variability in both quantitative and discrete traits (Darwin, 1859, 1868).

No progress in understanding the causes of this variability was made, however, until the rediscovery of Mendelian genetics at the beginning of the 20th century. The Mendelian basis for the inheritance of naturally occurring discrete polymorphisms, such as the ABO blood groups of humans (Bernstein, 1925), Batesian mimics in *Papilio* butterflies (Punnett, 1915) and heterostyly in *Primula* (Bateson & Gregory, 1905), was quickly established. However, discrete variation that is easily detectable at the phenotypic level is comparatively uncommon, except for rare deleterious mutations. In contrast, quantitative variation in meristic or metric traits is abundant, but less amenable to genetic analysis. By the 1920s, the joint control of quantitative of variation by non-genetic effects and multiple Mendelian genes was firmly established (Provine, 1971), and it could confidently be assumed that the vast bulk of heritable variation reflects the effects of underlying Mendelian variants carried on the chromosomes (Muller & Altenburg, 1920).

In addition, quantitative studies of the effects of inbreeding, such as those of Sewall Wright on guinea pigs (Wright, 1922), provided evidence for the existence of 'concealed variability', which is exposed when recessive alleles are made homozygous by matings between close relatives. The introduction of H. J. Muller's 'balancer' crossover suppressor chromosomes (Muller, 1928) into *Drosophila* population genetics in the 1930s, allowing the forced homozygosity of whole chromosomes derived from natural populations, yielded the startling conclusion that wild individuals of *Drosophila* species carry an average of over one recessive lethal mutation in the heterozygous state (Simmons & Crow, 1977). Ingenious competition experiments devised by John Sved later showed that homozygosity for a lethal-free

* Corresponding author. Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK. Tel: 44-(0)131-650-5750. Fax. 44-(0)131-650-6564. e-mail: Brian.Charlesworth@ed.ac.uk

haploid *D. melanogaster* genome reduces fitness to effectively zero under laboratory conditions (Sved, 1971; Latter & Sved, 1994).

While confirming Darwin's view that there is plenty of genetic variability available for use in evolution, the evidence obtained by these methods left two important questions unanswered (Lewontin, 1974). First, how much variation within a natural population is there at an average gene locus? Two radically different hypotheses had been proposed during the 1950s. Under the 'classical' view, most genes have a high-frequency, wild-type allelic form, *v.* some rare deleterious variants caused by mutation, like the recessive lethals described above (Muller, 1950). In contrast, the 'balance' hypothesis proposed that genes typically have several different allelic forms that segregate at intermediate frequencies, like the polymorphisms mentioned earlier (Dobzhansky, 1955). The second question concerned the extent to which the frequencies of variants within populations (other than rare deleterious mutations) are controlled by natural selection, as envisaged under the balance hypothesis, *v.* reflecting an interaction between mutation and random genetic drift, as proposed by the 'neo-classical' view (Kimura & Crow, 1964).

The methods of classical and quantitative genetics provide no means of sampling variation randomly from the genome, so that the first question cannot be answered by them. However, as shown below, modern DNA sequencing technology allows a virtually complete answer. The data provided by this technology also provide information that should allow the second question to be answered, but we are still some way from being confident that we know the answer. In this paper, I will give a brief historical review of how the first question has been answered, and then discuss methods that have been developed to answer the second question. Due to space constraints, only some of these can be described in any detail. In particular, I will not trace the development of the theory of the coalescent process, which has played a fundamentally important role in many of the modern methods of statistical testing and inference in population genomics; excellent reviews of coalescent theory are provided by Hein *et al.* (2005) and Wakeley (2008). Some questions that are raised by studies of DNA sequence variation and evolution will also be considered.

2. Molecular genetics to the rescue of population genetics

(i) *The pre-DNA era – gel electrophoresis of proteins*

The first estimates of genome-wide levels of variation in populations were obtained in 1966, by using the then recent discovery that most genes correspond to stretches of DNA that code for polypeptides. Detection of variation in the sequence of a polypeptide allows us to infer the existence of variation in the corresponding DNA sequence. John Hubby and Richard Lewontin applied this idea to samples from natural populations of *Drosophila pseudoobscura* (Hubby & Lewontin, 1966; Lewontin & Hubby, 1966), and Harry Harris independently applied it to humans (Harris, 1966). Both groups used gel electrophoresis to screen populations for variants that affect the migration rates of proteins on a gel exposed to an electric current. Many different soluble proteins controlled by independent genes were studied, mostly enzymes with well-understood metabolic roles. The proteins were chosen purely because they could be studied easily, with no bias with respect to any prior knowledge of their level of variability.

These papers also introduced ways of summarizing the results of genotyping numerous individuals at dozens of loci, by means of measures such as the proportion of polymorphic loci, P (i.e. loci with at least one minority variant with a frequency greater than a cut-off such as 1 or 5%), and the genic diversity, H (often referred to as the 'heterozygosity'), measured by the mean over the set of loci of the frequency with which two randomly sampled alleles at a locus differ in state. These pioneering studies resulted in hundreds of 'find 'em and grind 'em' surveys of natural variability (Lewontin, 1974, 1985). These surveys estimated that a large fraction (e.g. 43% in *D. pseudoobscura*, 28% in humans) of loci is usually polymorphic, and that H is of the order of a few percent (12% for *D. pseudoobscura* and 7% for humans). This apparently overthrew the classical view of genetic variability.

But these results had several limitations. Most importantly, only amino acid changes that affect the mobility of proteins in gels (mostly associated with charge changes) can be detected by electrophoresis; these probably represent only about one-third of the total possible mutational changes in the amino acid sequence of a protein (Lewontin, 1985). Additionally, only changes in DNA sequences that affect protein sequences can be observed.

Furthermore, attempts to use data on electrophoretic variability to test the neo-classical hypothesis that polymorphic variants are neutral (Kimura & Crow, 1964; Kimura & Ohta, 1971), *v.* the alternative that they are maintained by balancing selection (Lewontin & Hubby, 1966), proved frustratingly inconclusive. Lewontin (1974, p. 189) remarked that'

> 'For many years, population genetics was an immensely rich and powerful theory with virtually no suitable facts on which to operate. .... Quite suddenly the situation has changed … and facts in profusion have been poured into the hopper of this theory machine. And from the other end has issued – nothing. It is not that the machinery does not work, for a great clashing of gears is audible, if not deafening, but it somehow cannot transform into a finished product the great volume of raw material that has been provided. The

entire relationship between the theory and the facts needs to be reconsidered.'

By the end of the 1970s it was clear that studies of variation at the level of the DNA sequence itself would be needed to deal with these problems. In addition, theoretical population geneticists, starting with Warren Ewens (Ewens, 1972), had begun to produce theoretical models that predicted the properties of samples from a population, as opposed to modelling the population as a whole, greatly improving our ability to test hypotheses against data (for later developments, see Hein *et al.* 2005; Wakeley, 2008; Charlesworth & Charlesworth, 2010). Together, advances in both experimental techniques and theoretical methods have led to a much less pessimistic assessment of the prospects for solving the problem of the causes of variation.

(ii) *The beginning of the DNA revolution–restriction mapping*

It is probably hard for people brought up with the wonders of PCR amplification, automated Sanger sequencing, and now next-generation sequencing, to realize how difficult it was for geneticists to develop tools for studying variation at the DNA level, especially because relatively large amounts of cellular material were needed for molecular characterization of an individual before PCR amplification was available.

The first studies of DNA sequence variation in organelle and nuclear genes were done in the later 1970s, using restriction enzymes to detect variation at sites that could be cut by them (see reviews by Avise, 1983; Kazazian *et al.*, 1983; Nei, 1983). With nuclear genes, Southern blotting with probes derived from cloned genes was needed to pin down the region of interest, making it harder to obtain data.

D. melanogaster was especially amenable to this approach, because stocks of flies homozygous for any of the three major chromosomes could be made using balancer chromosomes (see Introduction section), providing plenty of material for isolating DNA from a single haploid genome. During the 1980s, Chuck Langley's laboratory was especially active in generating surveys of natural variability at different locations in the genome by this method, providing the first overview of genome-wide variation (Langley *et al.*, 1982, 1988; Aquadro *et al.*, 1986; Langley & Aquadro 1987; Schaeffer *et al.*, 1988; Miyashita & Langley, 1988; Aguadé, *et al.*, 1989b; Stephan & Langley, 1989; Aguadé *et al.*, 1992). Parallel studies were done in humans using cultured cells to provide material, starting with work on the human β-globin gene cluster (Kan & Dozy, 1978; Orkin *et al.*, 1982).

The analysis of data from restriction sites is not straightforward, since restriction maps need to be constructed for each haploid genome region to be characterized. Furthermore, since the only information provided is the presence or absence of the short sequences recognized by the enzymes, algorithms had to be developed to translate the restriction site variation observed in a sample into estimates of nucleotide site diversity (π), the per nucleotide site equivalent of H, which was introduced by Masatoshi Nei and Wen-Hsiung Li (Nei & Li, 1979) when analysing restriction site data; these are reviewed by Nei (1987, chapter 10). A considerable advantage of this approach was that a fairly large genomic region could be surveyed (13 kb in the case of the *D. melanogaster Adh* region and 50 kb in the case of the human β-globin gene cluster); in addition to changes between alternative nucleotides at a site, transposable element (TE) insertions, insertion/deletion (indel) polymorphisms and chromosome rearrangements could also be identified.

It is impressive how much information was gleaned from these early studies; for example, the results from both *Drosophila* and humans showed that what are now known as single nucleotide polymorphisms (SNPs) contributed the most to variability, in terms of events per nucleotide site, while TE insertions contributed low-frequency polymorphisms in *Drosophila* but almost nothing to human variability. It is interesting to note that Alan Robertson and Bill Hill (Robertson & Hill, 1983) used the results of Orkin *et al.* (1982) to infer that the mean nucleotide site diversity in humans implies an effective population size of 20 000, and that the disagreement between the observed level of linkage disequilibrium and the theoretical formula for its magnitude under drift and recombination suggested that 'crossing over is not homogeneous along the DNA sequence'. These findings were essentially confirmed by the results of much larger, more recent, studies of DNA sequence variability (see below). Furthermore, the *Drosophila* studies showed that regions of the genome with low recombination had unusually low levels of genetic variability (Aguadé *et al.*, 1989*a*; Stephan & Langley, 1989); more generally, a positive correlation was to be found between the local rate of recombination experienced by a gene, in terms of map units per unit of physical distance, and the level of genetic variability (Begun & Aquadro, 1992), a relationship that has also stood the test of time (see the Discussion).

(iii) *The rise of DNA sequencing*

With the invention of PCR amplification for isolating specific regions of DNA with the aid of sequence-specific primers, and with the introduction of sequencing machines, DNA sequencing of multiple copies of the same region or set of regions of the genome (resequencing, as it has come to be called)

became the method of choice for surveying DNA sequence variation. Before these methods became available, the first thorough study of variability by the use of sequencing was that of the *Adh* gene of *D. melanogaster* by Marty Kreitman (Kreitman, 1983), who applied the very laborious procedure of manual Maxam–Gilbert sequencing to stocks of 11 independently isolated chromosomes made homozygous by a balancer. (This was not a truly random sample from the population, as there were approximately equal numbers of the fast and slow electrophoretic alleles.)

A significant finding of this study, confirmed by later resequencing work, was that most of the variability involved silent changes that do not affect the protein sequence: the sequence differences were either in non-coding regions, or were coding sequence changes that did not affect the amino acid sequence (synonymous variants). Indeed, in Kreitman's *D. melanogaster Adh* gene survey, only one amino acid polymorphism was detected – the one previously known to cause the difference between the fast and slow alleles. About 39 amino acid variants would have been found if the same level of variability applied to both silent and non-synonymous variants (Kreitman, 1983). This agreed with the results of contemporary analyses of the molecular evolution of DNA sequences, which had shown a pattern of much slower rate of evolution per nucleotide site for non-synonymous compared to silent changes, e.g. Kimura (1983, chapter 4). These results led to the by-now familiar conclusion that the majority of new mutations that change the amino acid sequence have such large deleterious effects on fitness that they contribute little to either within-population variation or divergence between species, compared with silent changes to the DNA sequence.

The methods used for characterizing variability at the DNA sequence level have advanced steadily due to technical advances and continual reductions in costs; whereas surveys of more than a dozen or so genes were prohibitively expensive as recently as the year 2000, except for well-funded investigators of humans and medically important microbes, even financially hard-pressed *Drosophila* population geneticists have become been able to characterize samples of individuals for hundreds of roughly 500 bp sequences (the sizes of reads from an automated Sanger sequencer), from different sites around the genomes (e.g. Andolfatto, 2007; Hutter *et al*., 2007). While it is very expensive to scale this up to whole genomes, analyses of resequenced whole genomes of microbes such as yeast (Liti *et al*., 2009; Schacherer *et al*., 2009) and even *D. simulans* (Begun *et al*., 2007) have now been published.

These studies show that the overall levels of diversity at silent nucleotide sites within a species, which are the least likely to be influenced by differences in selective constraints, vary enormously among different taxa from a low of about 0·1 % for humans to a high of over 8 % for the seasquirt *Ciona intestinalis* (estimated by comparing the two haploid genomes of a single individual: Small *et al*., 2007): see Figure 1.10 of Charlesworth & Charlesworth (2010). Although levels of diversity per nucleotide site are usually small, even the small, gene-dense genome of the bacterium *Escherichia coli* has 4·2 million nucleotide sites, including around 900 000 silent or synonymous sites (Blattner *et al*., 1997). Given that the mean silent site diversity in *E. coli* is about 2 % (Charlesworth & Eyre-Walker, 2006), it can be estimated that there are likely to be about 82 900 silent SNPs among 100 *E. coli* genomes (Charlesworth & Charlesworth, 2010, p. 31). Several million SNPs (non-coding, synonymous and non-synonymous) are present genome-wide in the populations of most multicellular organisms, with their much larger genomes. There is thus a wealth of genetic variation in natural populations that was undreamt of in the days of electrophoresis, even disregarding the contribution of insertions, deletions and copy number variants, whose importance is becoming increasingly recognized (e.g. Iafrate *et al*., 2004; Emerson *et al*., 2008; Kidd *et al*., 2008).

With the introduction of high-throughput sequencing methods there will shortly be large-scale studies of both human populations and model organisms like *D. melanogaster* and *Arabidopsis thaliana*, where hundreds or thousands of independent genomes are being resequenced (see http://www.1000genomes.org; http://www.dpgp.org; http://www.hgsc.bcm.tmc.edu/project-species-i-Drosophila_genRefPanel.hgsc; http://1001genomes.org); a preliminary study of *D. melanogaster* has already been published (Sackton *et al*., 2009). We will soon have the finest-scale resolution that is possible. This throws the ball firmly into the court of theoretical population geneticists and statistical geneticists, to provide both a theoretical framework for interpreting any patterns discerned in the data, and quantitative tools for testing hypotheses against the data. If Lewontin's machinery does not work with this material, then we should give up! But the approaches reviewed below suggest that there are reasons to hope that the retooled machinery will perform rather well.

3. Inferring the causes of genome-wide variation

Simply estimating mean levels of sequence diversity across the genome tells us nothing about the forces involved in creating and maintaining it. We would like to know to what extent variability for different classes of variants (non-synonymous, synonymous and non-coding) can be accounted for by the neutral model, according to which genetic drift acts on

selectively equivalent or nearly equivalent types of variant (Kimura, 1983). If selection needs to be invoked, what kind of selection typically operates, and what is its intensity? Do other evolutionary forces, such as biased gene conversion (BGC) or meiotic drive play a significant part?

(i) *Mutation*

On almost any view of evolution other than a neo-Lamarckian one (currently, but in my opinion unconvincingly, being advocated by a vocal minority: e.g. Jablonka & Raz, 2009), both neutral and adaptive evolution depend on new mutations, genetic or epigenetic, that arise during the transmission of the genetic material from parent to offspring. The ability to characterize the sequences of whole genomes, or large portions of genomes, is revolutionizing our knowledge of the mutational process, both in terms of the rates of occurrence of mutations and the relative frequencies of different types of mutational change (reviewed by Lynch, 2010). These studies show that, despite low mutation rates per base pair per generation in the nuclear genomes of multicellular organisms (of the order of 10^{-9}–10^{-8}), the per-genome mutation rate per generation in short-lived species with relatively small genomes, such as *Drosophila*, is substantially higher than one new mutation per zygote per generation, and probably about 100 in humans, with their longer generation time and much larger genome. In addition, there is an almost universal bias in favour of mutational changes from GC base pairs to AT base pairs, compared with AT to GC, and for transitions over transversions. By far the most frequent type of mutation is represented by single nucleotide substitutions, followed by small insertion/deletions (indels).

While these conclusions are not radically new, they are now based solidly on direct evidence. This provides an essential underpinning for interpreting the results of population-level studies of variation. In addition, the results can be combined with estimates of the levels of selective constraints on amino-acid sequences and non-coding sequences, to yield estimates of the mean number of new deleterious mutations arising per individual per generation, U (Kondrashov & Crow, 1993; Haag-Liautard *et al*., 2007; Eory *et al*., 2010; Lynch 2010). This quantity plays a major role in theories of the evolution of genetic recombination, and the causes of inbreeding depression and ageing. It seems now well established from these studies that U is much larger than one for humans and about one for *Drosophila*. No doubt these estimates will be revised in the future with more data, but it seems as though we are close to settling a long-running dispute about the magnitude of U in higher organisms (for the background to this controversy, see Keightley & Eyre-Walker, 1999; Lynch *et al*., 1999).

(ii) *Selection on non-synonymous mutations*

(a) *The neutral null model*

Partly reflecting the fact that evolutionary and population studies of protein sequence differences have a much longer history than those of DNA sequences, a perhaps disproportionate amount of attention has been devoted to non-synonymous variation. It should be emphasized that the neutral theory as developed by Motoo Kimura and Tomoko Ohta (Kimura & Ohta, 1971; Kimura, 1983; Ohta, 1992) included the fact that most non-synonymous variants are sufficiently selectively deleterious that they have essentially no chance of fixation by genetic drift in opposition to selection, i.e. their $N_e s$ is usually substantially greater than 1, where N_e is the effective population size, and s is the selection coefficient against a deleterious non-synonymous mutation (measured in the heterozygous state with wild type, in the case of a diploid, randomly mating population).

The neutral theory proposes that the majority of non-synonymous variants that become fixed during evolution are the result of drift acting on neutral or nearly neutral mutations (i.e. those with $N_e s < 1$). Variants within populations are thus mainly either neutral or slightly deleterious, and are destined to ultimate fixation or loss by drift. Classical theoretical results on the neutral theory are as follows. The rate of sequence divergence between species per nucleotide site is equal to the corresponding mutation rate u. If $N_e u \ll 1$, the equilibrium level of nucleotide site diversity, π, under mutation and drift is equal to $4N_e u$. The probability that a polymorphic mutation is found at frequency q in a sample (the 'site frequency spectrum' or SFS) is proportional to $1/q$.

These classical findings provide the basis for many different tests of the neutral theory. A serious limitation, however, is that they assume that neutral sites are in statistical equilibrium under mutation and genetic drift in a panmictic population. While the assumption of panmixis is probably reasonably accurate for many species of *Drosophila*, it certainly does not apply to humans or predominantly self-fertilizing model organisms such as *C. elegans* and *Arabidopsis thaliana*. Furthermore, several intensively studied *Drosophila* populations show evidence for recent changes in population size, with bottlenecks and subsequent expansion (Haddrill *et al*., 2005*b*; Ometto *et al*., 2005), as is also the case for non-African human populations (Boyko *et al*., 2008). There has, therefore, been considerable effort to develop statistical tests for selection that include departures from the assumptions of the standard model, or even avoid them completely (see below).

(b) *Positive selection and the McDonald–Kreitman (MK) test*

There have been several approaches to testing the neutral model against the alternative hypothesis that many protein sequence variants are under positive selection, causing them to become fixed by selection on a time scale that is much faster than that of genetic drift. Perhaps the most successful has been the MK test (McDonald & Kreitman, 1991) and its extensions. Assume that the same length of sequence is used for both polymorphism and divergence estimates, so that the total numbers of mutable sites are the same in both cases. Under the null hypothesis that all types of sequence variants are neutral, the results quoted just above imply that the expected numbers of synonymous and non-synonymous differences in a coding sequence are proportional to the mutation rates for the two classes of variants: between-species differences and within-species polymorphisms.

Selection on the protein sequence produces a departure from this proportionality, which can be tested for by a simple 2×2 contingency table (but see Andolfatto, 2008). If some amino acid substitutions that distinguish a pair of species have been fixed by relatively strong directional ('positive') selection, they make little or no contribution to variation within a species, but increase the between-species divergence relative to its neutral expectation. The ratio of non-synonymous to synonymous between-species differences will then be elevated for the between-species comparison relative to the within-species diversity.

The MK test also provides a way of estimating the proportion of fixed differences between species at non-synonymous sites that were caused by positive selection, as opposed to genetic drift – it makes intuitive sense that the larger the ratio used in the MK test, the larger the proportion of non-synonymous differences that distinguish a pair of related species (commonly denoted by α). This intuition can be put into a more precise mathematical framework, and various methods for estimating α from MK tables for sets of loci have been developed (Fay *et al.*, 2002; Smith & Eyre-Walker, 2002; Welch, 2006).

Several different multi-locus surveys of DNA sequence variability in *Drosophila*, combined with estimates of divergence between two closely related species, have consistently suggested α values between 0·25 and 0·70, implying that a sizeable proportion of non-synonymous fixed differences are the result of positive selection (Eyre-Walker, 2006; Haddrill *et al.*, 2010). Similarly, data from multiple genome sequences of *Escherichia coli* and *Salmonella typhimurium/enterica* suggested an α of about 50% (Charlesworth & Eyre-Walker, 2006). In contrast, human polymorphism data and divergence data have yielded little evidence for positive selection by MK-based methods, (e.g. Zhang & Li, 2005). In *Drosophila*, there is evidence that certain categories of genes, especially those involved in male reproductive functions and immunity, may have unusually high rates of protein sequence evolution and high α values (Baines *et al.*, 2008; Obbard *et al.*, 2009). Future large-scale studies of genome-wide variability should lead to more estimates of α for different categories of genes in a wide variety of species (for a recent application to a flowering plant species, see Slotte *et al.*, 2010).

(c) *Testing for selective sweeps*

Another widely used approach applies the principle of the hitchhiking effect (Maynard Smith & Haigh, 1974), in order to detect evolutionarily recent selective events. A selectively favourable mutation that arises as a unique event, and then spreads to fixation, will cause closely linked variants (present on the chromosome in which it arose) to become fixed along with it, resulting in a 'selective sweep' (Berry *et al.*, 1991). This leads to a signature of reduced variation at linked neutral sites in a region surrounding the target of selection, provided that the ratio of their frequency of recombination with the site under selection, r, to the selective advantage of the mutation, s, is such that $r/s \ll 1$ (Maynard Smith & Haigh, 1974), and the sweep is sufficiently recent that variability has not been restored to its equilibrium level under drift and mutation. A sweep also causes a distortion of the SFS at sites near the target of selection, in favour of variants at extreme frequencies.

To detect such effects, a variety of tests have been proposed for evaluating the statistical significance of an observed reduction in variability, and the departure from the expected distribution of neutral variant frequencies in a sample from the population (the so-called site SFS), as well as other statistics such as levels of linkage disequilibrium (e.g. Hudson *et al.*, 1987; Braverman *et al.*, 1995; Simonsen *et al.*, 1995; Fay & Wu, 2000; Harr *et al.*, 2002; Kim & Stephan, 2002; Jensen *et al.*, 2005; Zeng *et al.*, 2007; Boitard *et al.*, 2009), or by comparing the typical genome-wide SFS with the SFS for candidates for selective sweeps, e.g. Nielsen *et al.* (2005). Several of these methods also include means of jointly estimating demographic changes and the locations of selective sweeps. These approaches have been successfully applied to the detection of sweeps, and to determine more or less precisely the location of the target of selection, especially in humans and *Drosophila* (for overviews, see Williamson *et al.*, 2007; Stephan 2010).

(d) *Testing for balancing selection*

If a new, selectively favourable mutation does not spread to fixation, but instead is subject to balancing

selection, or its selective advantage is restricted to certain local populations, then the haplotype associated with the new variant will initially be present at an intermediate frequency, with a high level of linkage disequilibrium with respect to variants at surrounding sites. Before recombination has had the opportunity to introduce variants from haplotypes that lack the favoured variant, those that carry it will therefore show a low level of genetic diversity at linked sites. As time goes on, recombination whittles this effect away, so the magnitude of reduction in diversity among haplotypes carrying the new variant depends jointly on the time since it originated and the rate of recombination. Several methods have been developed for assessing the statistical significance of a region of 'extended homozygosity', associated with the relatively recent spread of a selectively favoured variant to an intermediate frequency; these methods have also been successfully used to locate the targets of selection (e.g. Hudson *et al.*, 1994; Voight *et al.*, 2005; Sabeti *et al.*, 2007; Coop *et al.*, 2009).

If two variants at a site, A_1 and A_2, are maintained by balancing selection for a long period of time, substantially greater than their expected coalescence time under neutrality ($2N_e$ generations), then a rather different pattern of variability at linked neutral sites is expected to develop. The flow of variants at a neutral site by recombination between chromosomes carrying A_1 and A_2 is similar to migration between different demes, and takes place at a rate proportional to r, the recombination frequency between the neutral and selected sites. Eventually, drift, mutation and recombination will come into equilibrium, in a way similar to that for a spatially rather than genetically subdivided population. High equilibrium levels of differentiation between the A_1 and A_2 haplotypes would thus be expected only at closely linked neutral sites (i.e., in the situation equivalent to low migration). This produces a local peak in neutral diversity around the target of selection, which decline over a genetic distance of the order $r = 1/N_e$; the SFS for variants close to the target of selection is distorted in favour of intermediate frequency variants (Hudson & Kaplan, 1988; Hudson, 1990; Nordborg, 1997; Navarro & Barton, 2002; Barton & Etheridge, 2004).

Such signatures of long-term balancing selection can be seen in the classic cases of the mammalian MHC genes (Shiina *et al.*, 2006) and the self-incompatibility (SI) genes of plants (Kamau *et al.*, 2007). They are sometimes incorrectly referred to as 'hitchhiking effects' (e.g. Shiina *et al.*, 2006), but it should be clear from the above description that no hitchhiking is involved, in the sense of changes in frequencies of neutral variants associated with the selectively driven change in frequency of a linked variant. However, genome-wide scans of human populations suggest that there are few cases of long-term balancing selection of this kind, although some exceptions have been detected that represent less than 1 % of genes surveyed (Bubb *et al.*, 2006; Andres *et al.*, 2009).

(e) *Testing for local selection*

Somewhat similar principles apply to the detection of differences between populations. Both the recent spread of a mutation that has failed to go to fixation in all local populations of a species (Slatkin & Wiehe, 1998), and the long-term maintenance of variants that are subject to opposing pressures in different locations (Charlesworth *et al.*, 1997), will produce an excess divergence between populations at linked neutral sites, compared to the genome-wide average. Systematic surveys for effects of this kind have been carried out in human populations, for example, and a large number of candidate genes involved in such selective differentiation between populations have been detected (Akey, 2009; Coop *et al.*, 2009; Hancock *et al.*, 2010).

(f) *Estimation of the distribution of mutational effects on fitness*

A different question about selection involves the nature of the probability distribution of the selection coefficients against new amino-acid mutations. This problem is of considerable importance for numerous issues in evolutionary genetics, including the extent to which selection against deleterious mutations affects evolution at closely linked neutral or weakly selected sites (the process of 'background selection': Charlesworth *et al.*, 1993), and the evolutionary significance of genetic recombination (Barton, 2010).

Two main methods have been developed for estimating the parameters of this distribution. One involves the comparison of the SFSs for non-synonymous variants and putatively neutral variants, such as synonymous or intron variants, fitting these to models that allow for a distribution, $\phi(s)$, of selection coefficients against heterozygous non-synonymous new mutations. These methods assume distributions with two parameters, such as the normal, log−normal or gamma distributions, and some of them correct for the effects of demographic changes on the SFSs (Piganeau & Eyre-Walker, 2003; Eyre-Walker *et al.*, 2006; Sawyer *et al.*, 2007; Keightley & Eyre-Walker, 2007; Boyko *et al.*, 2008). The other method relies on a comparison between two species with very different effective population sizes, as indicated by their levels of synonymous variability, which are assumed to be close to neutrality. The extent to which they also differ in their levels of non-synonymous variability reflects the nature of $\phi(s)$ (Loewe *et al.*, 2006; Haddrill *et al.*, 2010).

The results of these studies of both human and *Drosophila* populations suggest a wide and highly skewed distribution of selection coefficients of s, with most mutations being very weakly selected but with a long tail of much more strongly selected mutations, some even being effectively lethal. The mean of s is hard to estimate, but seems likely to be of the order of a few percent in the case of humans, and possibly an order of magnitude less for *Drosophila*. The mean selection coefficient against segregating mutations is, however, relatively small, so that their mean $N_e s$ of the order of 10 or less. A relatively small proportion of new mutations seem to be nearly neutral, with less than 10% having $N_e s < 0{\cdot}5$. Mutations more strongly selected than this behave essentially deterministically, as far as their level of nucleotide site diversity is concerned (McVean & Charlesworth, 1999).

Knowledge of $\phi(s)$ also allows estimation of the overall probability of fixation of a new non-synonymous variant, enabling predictions to be made of the expected amount of non-synonymous divergence between species (Loewe *et al.*, 2006; Boyko *et al.*, 2008; Eyre-Walker & Keightley, 2009), and hence of the value of α, by a different approach from the MK test (see section 3(ii)(b)). Implementation of this method in *Drosophila* and humans has yielded similar estimates to those from the MK test, with α in the region of 50% for *Drosophila*, but only 10% in humans (Boyko *et al.*, 2008; Eyre-Walker & Keightley, 2009; Haddrill *et al.*, 2010).

(iii) *Selection on non-coding variants*

It was assumed for some time by molecular population geneticists that most of the conveniently studied non-coding sequences, such as introns and untranslated transcribed regions (UTRs), would be under weaker selection than even synonymous sites, so that introns, for example, would provide a useful neutral proxy for use in MK tests, and in estimating demographic parameters.

A number of surveys of *Drosophila* populations using sets of approximately 500 bp intron sequences were conducted for the latter purpose (Haddrill *et al.*, 2005*b*; Ometto, *et al.*, 2005; Hutter, *et al.*, 2007). It quickly became apparent by comparing their interspecies sequence divergence with that of synonymous sites that such large (for *Drosophila*) intron sequences are under substantially higher levels of selective constraints than synonymous sites (Haddrill *et al.*, 2005*a*, *b*; Halligan & Keightley, 2006).

This has caused attention to be shifted to estimating the extent of positive and purifying selection on non-coding sequences. Many of the tests for selection on non-synonymous mutations can be applied to different types of non-coding sequences, and no major new principles are needed for this purpose. Overall, it seems that sites in short introns (length <100 bp or so) in *Drosophila* are nearly neutral (Parsch *et al.*, 2010), whereas sites in long introns are subject to predominantly purifying selection, and UTRs are subject to both purifying and positive selection (Andolfatto, 2005; Haddrill *et al.*, 2008). Similar studies of human non-coding sequences also provide evidence for both purifying and positive selection (e.g. The Encode Project Consortium, 2007).

However, when studying evolution and variation in non-coding sequences it is important to take into account the phenomenon of BGC. This refers to an excess over 50% of the frequency of one of the two variants at a heterozygous nucleotide site among the products of meiosis, associated with the formation of heteroduplex DNA (Marais, 2003). This is often associated with heterozygosity for a GC base pair and an AT base pair, with a bias in the direction of an excess of the GC variant, in which case the process is sometimes referred to as gBGC (the g refers to the G in GC). The net effect is similar to positive selection (Gutz & Leslie, 1976), so that gBGC causes a higher probability of fixation of the GC variant relative to neutrality, and a lower probability of fixation of the AT variant. Selection on GC versus AT variants is thus confounded with the effect of gBGC (Galtier & Duret, 2007).

(iv) *Selection on synonymous variants*

The fact that synonymous mutations are under weaker selective constraints than many types of non-coding sequences does not imply that they are completely neutral. Since the 1980s, evidence has accumulated that codon usage bias in many species reflects the action of natural selection, most probably involving translational efficiency or accuracy (Ikemura, 1982; Sharp & Li, 1986; Drummond & Wilke, 2008; Sharp *et al.*, 2010). A variety of methods have been developed for using data on the population frequencies of alternative synonymous variants, which correspond to codons that have been identified from codon usage studies as 'preferred' versus 'non-preferred' (i.e. codons that are over- versus under-represented in genes with biased codon usage). If large numbers of polymorphic synonymous sites are available, model fits can be applied to estimate the intensity of selection, usually expressed in terms of the product of N_e and the selection coefficient in favour of a preferred synonymous variant at a site (Hartl *et al.*, 1994; Akashi, 1995, 1999; Maside *et al.*, 2004; Comeron & Guthrie, 2005; Cutter & Charlesworth, 2006).

The most recently developed methods also include fits to models of population size changes, which can significantly bias estimates of selection if they are ignored (Zeng & Charlesworth, 2009, 2010; Zeng,

2010). Conversely, apparent evidence for population expansion that is obtained under the assumption of neutrality at synonymous sites may disappear if selection is taken into account, as in the case of the Zimbabwe population of *D. melanogaster* (Zeng & Charlesworth, 2009). In several *Drosophila* species, evidence for $N_e s$ values for preferred versus unpreferred codons in the region of 0·5 has been obtained by these methods. Given that the effective sizes of the species concerned are in the millions, selection coefficients of the order of 10^{-7} to 10^{-6} are being detected, which of course is far below the resolution of any experimental approach.

With sufficiently large data sets, it has been possible to examine the relationships between $N_e s$ estimates obtained in this way, and factors such as coding sequence length and gene expression levels, which are known to be related to codon usage. In *Drosophila*, these studies have shown that $N_e s$ is significantly lower in longer genes than shorter ones, in genes with low expression rather than high expression, and in the middle of genes compared to their ends, and on the *X* chromosome compared with the autosomes (Comeron & Guthrie, 2005; Zeng & Charlesworth, 2009, 2010). Evidence is accumulating that selection also acts on synonymous sites in human populations (Comeron, 2006; Kondrashov *et al*., 2006), although factors other than translational efficiency are likely to be involved, notably on selection on mutations affecting exon splice sites and nucleosome positioning (Parmley & Hurst, 2007). With the advent of genome-wide resequencing data, it should shortly prove possible to greatly extend these types of analyses.

The possible role of BGC (see section 3(iii) above) in creating apparent selection on non-coding and synonymous sites can also be investigated by population genetics methods applied to large data sets. Since preferred codons in *Drosophila* and *E. coli* mostly end in GC, and most synonymous changes involve third coding positions, the effects of gBGC and selection on codon usage on synonymous sites may be confounded to a considerable extent, so that estimates of $N_e s$ will reflect both forces. By estimating the intensity of apparent selection in favour of GC over AT base pairs at intron sites in the genes that are also used for estimating $N_e s$ for synonymous sites, the true intensity of selection on synonymous sites can in principle be estimated. Studies of this kind in *Drosophila* suggest that the equivalent of $N_e s$ for gBGC ($N_e \omega$) is on average about one-quarter of the typical values for synonymous sites (Zeng & Charlesworth, 2010), so that most of the apparent selection on the latter probably reflects selection in favour of preferred codons.

Nevertheless, gBGC appears to be a significant factor in affecting the base composition of the genome, especially at non-coding sites, and may even affect evolution at non-synonymous sites in 'hotspots' of unusually high recombination frequencies (Ratnakumar *et al*., 2010). In particular, regions of the genome with high GC content seem to have higher $N_e \omega$ values than regions with lower GC content in both humans (Duret & Arndt, 2008) and *Drosophila* (Galtier *et al*., 2006; Haddrill & Charlesworth, 2008). There is, however, a paradox: if $N_e \omega$ is much smaller than $N_e s$ for preferred codons, why do non-coding sequences in *Drosophila* often show much larger selective constraints than synonymous sites (see section 3(iii) above)? The answer must lie in selective pressures on the sequences of non-coding sequences that are unrelated to selection or gBGC; a recent analysis of polymorphism data in *D. melanogaster* has provided evidence for such an effect (Zeng & Charlesworth, 2010).

4. Conclusions and broader implications

The advent of resequencing studies of whole genomes will greatly increase the amount of data available, and the power of methods of inference concerning the forces involved in DNA sequence variation and evolution. Almost certainly, it will stimulate the development of ever-more sophisticated and computationally demanding methods of data interpretation. It is thus likely that many details of the results described above will be substantially revised in the not-too-distant future. Nonetheless, I am optimistic that their broad outlines will turn out to be roughly correct.

What, then, have we discovered as a result of these efforts, compared say with the state of the field in 1983, when Kimura summed up his views on molecular evolution and variation (Kimura 1983)? One feature that seems to stand out is that the neutral theory is now regarded mainly as a null model, against which alternatives such as selection and BGC can be tested. We now have fairly solid evidence that a substantial fraction of non-synonymous differences between species, and of certain types of non-coding differences, have been caused by positive selection, at least for species with large effective population sizes such as bacteria, *Drosophila* and some plants, but less so for humans. This may reflect a greater contribution in humans from the fixation by drift of slightly deleterious mutations, due to their low effective population size (Eyre-Walker *et al*., 2002), and does not necessarily imply an overall lower rate of fixation per gene of favourable mutations.

We also have some confidence that most new non-synonymous mutations are sufficiently strongly selected against that they have little chance of fixation by drift; nevertheless, the mean selection coefficient against an amino-acid mutation that is segregating in a population is very small, of the order of 10^{-3} or less in the case of humans. This generates the perhaps

startling conclusion that individuals in populations of outbred organisms typically carry large numbers of deleterious amino acid variants that are effectively maintained by a balance between mutation and selection, of the order of 800 for humans (Eyre-Walker *et al*., 2006; Kryukov *et al*., 2007) and 4500 for *Drosophila* (Haddrill *et al*., 2010). As several authors have pointed out (Pritchard, 2001; Wright *et al*., 2003; Eyre-Walker 2006, 2010; Kryukov *et al*., 2007), the presence of such a large number of low-frequency deleterious mutations in populations creates a substantial variance in fitness and in the traits that they influence, even if the fitness effects of individual mutations are small. The existence of such a variance has long been inferred from studies of mutational effects, concealed variation and the genetic variance in fitness components of *Drosophila* (Simmons & Crow, 1977; Lynch *et al*., 1999; Charlesworth & Hughes, 2000), but this evidence was largely overlooked by the human genetics community.

This suggests that much human genetic susceptibility to disease may be the effects of rare mutations with minor phenotypic effects. It follows that even very large-scale searches for associations between genetic markers and diseases may uncover only the (possibly small) portion of the total variability caused by common major effect variants. There is indeed increasing evidence that many common diseases with a strong genetic component are caused by larger numbers of low-frequency variants, both non-coding and non-synonymous (Kryukov *et al*., 2007; Eyre-Walker, 2010). Ironically, therefore, the classical view of the maintenance of genetic variation affecting fitness-related traits (Muller, 1950) has been partially vindicated.

The prevalence of so much selection across the genome raises many questions. I will only discuss two. The first is the classic one of how a species with a large genome and a relatively low maximal reproductive rate, such as humans, withstands the resulting very high genetic load arising from the constant input of deleterious mutations, at a rate substantially greater than one new mutation per generation (Muller, 1950; Crow, 1997). As Alexey Kondrashov once put it, why have we not died 100 times over (Kondrashov, 1995)? It remains to be determined how serious this problem actually is, once we get a better idea of how much of the non-coding DNA is under selection. If it is as serious as seems likely to be the case, then Kondrashov's proposal that it can only be resolved by some form of quasi-truncation selection needs to be thoroughly examined. The related question of the long-term consequences of relaxing selection against deleterious mutations by medical intervention against human genetic disease also needs attention (Crow, 1997; Lynch, 2010).

The second question concerns the extent to which selection, both purifying and positive, at a multiplicity of sites across the genome has effects on variation and adaptation at nearby sites, as a result of Hill–Robertson (HR) interference (Hill & Robertson, 1966). Selection creates heritable variance in fitness among individuals, which reduces N_e. A site that is linked to a selected variant experiences an especially marked reduction in its N_e, because close linkage maintains the effects for many generations (Hill & Robertson, 1966; Comeron *et al*., 2008; Barton, 2010). In addition to reducing levels of variability, this reduction in N_e impairs the efficacy of selection, since the chance of fixation of a mutation depends on the product of N_e and its selection coefficient. Both selective sweeps and background selection constitute forms of HR interference, which almost certainly accounts for the correlation between the local recombination rate and the level of silent sequence diversity in *Drosophila* noted in section 2(ii) (see also Presgraves, 2005; Shapiro *et al*., 2007), since interference is less likely when recombination rates are high. There is increasing evidence for such an effect in other taxa, including humans and *Caenorhabditis* (Cai *et al*., 2009; Cutter & Choi, 2010; Rockman *et al*., 2010). Similarly, the sharply reduced level of codon usage and the accelerated rate of protein sequence evolution due to relaxed purifying selection, which are observed in low recombination regions of the *Drosophila* genome, are consistent with HR interference (Arguello *et al*., 2010; Charlesworth *et al*., 2010).

Almost certainly, patterns of evolution and variation in organisms with low effective recombination rates, such as bacteria with their limited rates of genetic exchange among individuals, or highly homozygous selfing species such as budding yeasts and *C. elegans*, will be subject to strong HR effects, yet this has scarcely been explored in studies of their population genomics. There is also an important question about the effect of HR interference in regions of the genome with 'normal' rates of recombination in outbred species. There are empirical indications that such effects exist, such as the negative relation between the rate of protein sequence evolution of a gene and its level of silent site diversity in *Drosophila* (Sella *et al*., 2009), reduced diversity near coding sequences in humans (Cai *et al*., 2009; McVicker *et al*., 2009; Hammer *et al*., 2010), and the fact that codon usage bias in *D. melanogaster* is lower in the middle of genes and in genes that lack introns (Comeron & Kreitman, 2002). The first of these observations is most easily explained by selective sweeps, the others may well involve background selection effects (Loewe & Charlesworth, 2007; McVicker *et al*., 2009). In either case, it is seems clear that a full understanding of patterns revealed by genome-level studies will involve the inclusion of the joint effects of

selection and linkage, which have so far largely been ignored in the modelling machinery used for inference.

References

Aguadé, M., Miyashita, N. & Langley, C. H. (1989*a*). Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. *Genetics* **122**, 607–615.

Aguadé, M., Miyashita, N. & Langley, C. H. (1989*b*). Restriction-map variation at the *zeste-tko* region in natural populations of *Drosophila melanogaster*. *Molecular Biology and Evolution* **6**, 123–130.

Aguadé, M., Miyashita, N. & Langley, C. H. (1992). Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila melanogaster*. *Genetics* **132**, 755–770.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in Drosophila DNA. *Genetics* **139**, 1067–1076.

Akashi, H. (1999). Inferring the fitness effects of DNA polymorphisms and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**, 221–238.

Akey, J. M. (2009). Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research* **19**, 711–722.

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149–1152.

Andolfatto, P. (2007). Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Research* **17**, 1755–1762.

Andolfatto, P. (2008). Controlling type-I error of the McDonald–Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* **180**, 1767–1771.

Andres, A. M., Hubisz, M. J., Indap, A., Torgerson, D. G., Degenhardt, J. D., Boyko, A. R., Gutenkunst, R. N., White, T. J., Green, E. D., Bustamante, C. D., Clark, A. G., Nielsen, R. (2009). Targets of balancing selection in the human genome. *Molecular Biology and Evolution* **26**, 2755–2764.

Aquadro, C. F., Deese, S. F., Bland, M. M., Langley, C. H. & Laurie-Ahlberg, C. C. (1986). Molecular population genetics of the *Alcohol dehydrogenase* gene region of *Drosophila melanogaster*. *Genetics* **114**, 1165–1190.

Arguello, J. R., Zhang, Y., Kado, T., Fan, C. Z., Zhao, R. P., Innan, H., Wang W. Long, M. Y. (2010). Recombination yet inefficient selection along the *Drosophila melanogaster* subgroup's fourth chromosome. *Molecular Biology and Evolution* **27**, 848–861.

Avise, J. C. (1983). Polymorphism of mitochondrial DNA in populations of higher animals. In *Evolution of Genes and Proteins* (ed. M. Nei & R. K. Koehn), pp. 147–164. Sunderland, MA: Sinauer Associates.

Baines, J. F., Sawyer, S. A., Hartl, D. L. & Parsch, J. (2008). Effects of sex-linkage and sex-biased gene expression in the rate of adaptive protein evolution in *Drosophila*. *Molecular Biology and Evolution* **25**, 1639–1650.

Barton, N. H. (2010). Genetic linkage and natural selection. *Philosophical Transactions of the Royal Society B* **365**, 2559–2569.

Barton, N. H. & Etheridge, A. M. (2004). The effect of selection on genealogies. *Genetics* **166**, 1115–1131.

Bateson, W. & Gregory, R. P. (1905). On the inheritance of heterostylism in Primula. *Proceedings of the Royal Society B* **76**, 581–586.

Begun, D. J. & Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. *Nature* **356**, 519–520.

Bengun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. P., Nista, P. M., Jones, C. B., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *Public Library of Science Biology* **5**, e310.

Bernstein, F. (1925). Zusammenfassende Betrachtungen aus der Theorie der Blutgruppen. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **37**, 237–269.

Berry, A. J., Ajioka, J. W. & Kreitman, M. (1991). Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129**, 1111–1117.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., ColladoVides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.

Boitard, S., Schlötterer, C. & Futschik, A. (2009). Detecting selective sweeps: a new approach based on hidden Markov models. *Genetics* **181**, 1567–1578.

Boyko, A., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., Bustamante, C. D. (2008). Assessing the evolutionary impact of amino-acid mutations in the human genome. *Public Library of Science Genetics* **5**, e1000083.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* **140**, 783–796.

Bubb, K. L., Bovee, D., Buckley, D., Haugen, E., Kibukawa, M., Paddock, M., Palmieri, A., Subramanian, S., Zhou, Y., Kaul, R., Green, P., Olson, M. V. (2006). Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**, 2165–2177.

Cai, J. J., Macpherson, J. M., Sella, G. & Petrov, D. A. (2009). Pervasive hitchhiking at coding and regulatory sites in humans. *Public Library of Science Genetics* **5**, e1000336.

Charlesworth, B., Betancourt, A. J., Kaiser, V. B. & Gordo, I. (2010). Genetic recombination and molecular evolution. *Cold Spring Harbor Symposia on Quantitative Biology* **74**, 177–186.

Charlesworth, B. & Charlesworth, D. (2010). *Elements of Evolutionary Genetics*. Greenwood Village, CO: Roberts and Company.

Charlesworth, B. & Hughes, K. A. (2000). The maintenance of genetic variation in life-history traits. In *Evolutionary Genetics from Molecules to Morphology* (ed. R. S. Singh & C. B. Krimbas), pp. 369–392. Cambridge: Cambridge University Press.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Charlesworth, B., Nordborg, M. & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**, 155–174.

Charlesworth, J. & Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution* **23**, 1348–1356.

Comeron, J. M. (2006). Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proceedings of the National Academy of Sciences of the USA*. **103**, 6940–6945.

Comeron, J. M. & Guthrie, T. B. (2005). Intragenic Hill–Robertson interference influences selection on synonymous mutations in *Drosophila*. *Molecular Biology and Evolution* **22**, 2519–2530.

Comeron, J. M. & Kreitman, M. (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics* **161**, 389–410.

Comeron, J. M., Williford, A. & Kliman, R. M. (2008). The Hill–Robertson effect: evolutionary consequences of weak selection in finite populations. *Heredity* **100**, 19–31.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W., Pritchard, J. K. (2009). The role of geography in human evolution. *Public Library of Science Genetics* **5**, e1000500.

Crow, J. F. (1997). The high spontaneous mutation rate: is it a health risk? *Proceedings of the National Academy of Sciences of the USA*. **94**, 8380–8386.

Cutter, A. & Choi, J. Y. (2010). Natural selection shapes nucleotide polymorphism across the genome of the nematode *Caenorhabditis briggsae*. *Genome Research* **20**, 1103–1111.

Cutter, A. D. & Charlesworth, B. (2006). Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabitis remanei*. *Current Biology* **16**, 2053–2057.

Darwin, C. R. (1859). *The Origin of Species*. London: John Murray.

Darwin, C. R. (1868). *The Variation of Animals and Plants under Domestication*. 2 Vols. London: John Murray.

Dobzhansky, T. (1955). A review of some fundamental concepts and problems of population genetics. *Cold Spring Harbor Symposia on Quantitative Biology* **20**, 1–15.

Drummond, D. A. & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **13**, 341–352.

Duret, L. & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *Public Library of Science Genetics* **4**, e10000071.

Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. & Long, M. (2008). Natural selection shapes genome wide patterns of copy number polymorphism in *D. melanogaster*. *Science* **320**, 1629–1631.

Eory, L., Halligan, D. L. & Keightley, P. D. (2010). Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Molecular Biology and Evolution* **27**, 177–192.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in Ecology and Evolution* **21**, 569–575.

Eyre-Walker, A. (2010). Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences of the USA*. **107**, 1752–1756.

Eyre-Walker, A. & Keightley, P. D. (2009). Estimating the rate of adaptive mutations in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution* **26**, 2097–2108.

Eyre-Walker, A., Keightley, P. D., Smith, N. G. C. & Gaffney, D. (2002). Quantifying the slightly deleterious mutation model of molecular evolution. *Molecular Biology and Evolution* **19**, 2142–2149.

Eyre-Walker, A., Woolfit, M. & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino-acid mutations in humans. *Genetics* **173**, 891–900.

Fay, J., Wykhoff, G. J. & Wu, C.-I. (2002). Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026.

Fay, J. C. & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413.

Galtier, N., Bazin, E. & Bierne, N. (2006). GC-biased segregation of non-coding polymorphisms in *Drosophila*. *Genetics* **172**, 221–228.

Galtier, N. & Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics* **23**, 273–277.

Gutz, H. & Leslie, J. F. (1976). Gene conversion: a hitherto overlooked parameter in population genetics. *Genetics* **83**, 861–866.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Houle, D., Charlesworth, B., Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445**, 82–85.

Haddrill, P. R., Bachtrog, D. & Andolfatto, P. (2008). Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Molecular Biology and Evolution* **25**, 1825–1834.

Haddrill, P. R. & Charlesworth, B. (2008). Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biology Letters* **4**, 438–441.

Haddrill, P. R., Charlesworth, B., Halligan, D. L. & Andolfatto, P. (2005*a*). Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biology* **6**, R67.1–R67.8.

Haddrill, P. R., Loewe, L. & Charlesworth, B. (2010). Estimating the parameters of selection on nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. *Genetics* **185**, 1381–1396.

Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. (2005*b*). Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research* **15**, 790–799.

Halligan, D. L. & Keightley, P. D. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide sequence comparison. *Genome Research* **16**, 875–884.

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Cox, M. P. & Wall, J. D. (2010). The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nature Genetics* **41**, 830–831.

Hancock, A. M., Alkorta-Aranbururu, G., Witonsky, D. B. & Di Rienzo, A. (2010). Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society B* **365**, 2459–2468.

Harr, B., Kauer, M. & Schloetterer, C. (2002). Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the USA*. **99**, 12949–12954.

Harris, H. (1966). Enzyme polymorphisms in man. *Proceedings of the Royal Society B*. **164**, 298–310.

Hartl, D. L., Moriyama, E. N. & Sawyer, S. A. (1994). Selection intensity for codon bias. *Genetics* **138**, 227–234.

Hein, J., Schierup, M. H. & Wiuf, C. 2005. *Gene Genealogies, Variation and Evolution*. Oxford: Oxford University Press.

Hill, W. G. & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–294.

Hubby, J. L. & Lewontin, R. C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**, 577–594.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys of Evolutionary Biology* **7**, 1–45.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. (1994). Evidence for positive selection in the *Superoxide dismutase* (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**, 1329–1340.

Hudson, R. R. & Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics* **120**, 831–840.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987). A test of molecular evolution based on nucleotide data. *Genetics* **116**, 153–159.

Hutter, S., Li, H. P., Beisswanger, S., De Lorenzo, D. & Stephan, W. (2007). Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosome-wide nucleotide polymorphism data. *Genetics* **177**, 469–480.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listwenik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes; differences in synonymous codon choice patterns of yeast and *E. coli* with reference to the abundance of isoaccepting transfer RNAs. *Journal of Molecular Evolution* **158**, 389–409.

Jablonka, E. & Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence mechanisms, and implications for the study of heredity and evolution. *Quarterly Review of Biology* **84**, 131–176.

Jensen, J. D., Kim, Y., Bauer DuMont, V., Aquadro, C. F. & Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401–1410.

Kamau, E., Charlesworth, B. & Charlesworth, D. (2007). Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of *Arabidopsis lyrata*. *Genetics* **176**, 2357–2369.

Kan, Y. W. & Dozy, A. M. (1978). Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relation to sickle mutation. *Proceedings of the National Academy of Sciences of the USA* **75**, 5631–5635.

Kazazian, H., Chakravarti, A., Orkin, S. H. & Antonarakis, S. E. (1983). DNA Polymorphisms in the human β globin gene cluster. In *Evolution of Genes and Proteins* (ed. M. Nei & R. K. Koehn), pp. 137–146. Sunderland, MA: Sinauer.

Keightley, P. D. & Eyre-Walker, A. (1999). Terumi Mukai and the riddle of deleterious mutation rates. *Genetics* **153**, 515–523.

Keightley, P. D. & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**, 2251–2261.

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64.

Kim, Y. & Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kimura, M. & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

Kimura, M. & Ohta, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton, NJ: Princeton University Press.

Kondrashov, A. S. (1995). Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of Theoretical Biology* **175**, 583–594.

Kondrashov, A. S. & Crow, J. F. (1993). A molecular approach to estimating the human deleterious mutation rate. *Human Mutation* **2**, 229–234.

Kondrashov, F. A., Ogurtsov, A. Y. & Kondrashov, A. S. (2006). Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *Journal of Theoretical Biology* **240**, 616–626.

Kreitman, M. (1983). Nucleotide polymorphism at the *Alcohol dehydrogenase* locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Kryukov, G. V., Pennachio, L. A. & Sunyaev, S. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *American Journal of Human Genetics* **80**, 727–739.

Langley, C. H., Aquadro, C. F. (1987). Restriction map variation in natural populations of *Drosophila melanogaster*: *white* locus region. *Molecular Biology and Evolution* **4**, 651–663.

Langley, C. H., Montgomery, E. A. & Quattlebaum, W. F. (1982). Restriction map variation in the Adh region of Drosophila. *Proceedings of the National Academy of Sciences of the USA* **79**, 5631–5635.

Langley, C. H., Shrimpton, A. E., Yamazaki, T., Miyashita, N., Matsuo, Y. & Aquadro, C. F. (1988). Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. *Genetics* **119**, 619–629.

Latter, B. D. H. & Sved, J. A. (1994). A re-evaluation of data from competitive tests shows high levels of heterosis in *Drosophila melanogaster*. *Genetics* **137**, 509–511.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York, NY: Columbia University Press.

Lewontin, R. C. (1985). Population genetics. *Annual Review of Genetics* **19**, 81–102.

Lewontin, R. C. & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in *Drosophila pseudoobscura*. *Genetics* **54**, 595–609.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Durbin, R., Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341.

Loewe, L. & Charlesworth, B. (2007). Background selection in single genes may explain patterns of codon bias. *Genetics* **175**, 1381–1393.

Loewe, L., Charlesworth, B., Bartolomé, C. & Nöel, V. (2006). Estimating selection on nonsynonymous mutations. *Genetics* **172**, 1079–1092.

Lynch, M. (2010). Rate, molecular spectrum and consequences of human mutation. *Proceedings of the National Academy of Sciences of the USA*. **107**, 961–968.

Lynch, M., Blanchard, J., Houle, D., Kibota, T., Schultz, S., Vassilieva, L. & Willis, J. (1999). Perspective: spontaneous deleterious mutation. *Evolution* **53**, 645–663.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics* **19**, 330–338.
Maside, X., Weishan Lee, A. & Charlesworth, B. (2004). Selection on codon usage in *Drosophila americana*. *Current Biology* **14**, 150–154.
Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.
McDonald, J. H. & Kreitman, M. (1991). Accelerated protein evolution at the *Adh* locus in Drosophila. *Nature* **351**, 652–654.
McVean, G. A. T. & Charlesworth, B. (1999). A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genetical Research* **74**, 145–158.
McVicker, G., Gordon, D., Davis, C. & Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *Public Library of Science Genetics* **5**, e1000471.
Miyashita, N. & Langley, C. H. (1988). Molecular and phenotypic evolution of the *white* locus in *Drosophila melanogaster*. *Genetics* **120**, 199–212.
Muller, H. J. (1928). The measurement of mutation rate in Drosophila, its high variability and its dependence on temperature. *Genetics* 279–357.
Muller, H. J. (1950). Our load of mutations. *American Journal of Human Genetics* **2**, 111–176.
Muller, H. J. & Altenburg, E. (1920). The genetic basis of truncate wing – an inconstant and modifiable character in Drosophila. *Genetics* **5**, 1–59.
Navarro, A. & Barton, N. H. (2002). The effects of multilocus balancing selection on neutral variability. *Genetics* **161**, 849–863.
Nei, M. (1983). Genetic polymorphism and the role of mutation in evolution. In *Evolution of Genes and Proteins* (ed. M. Nei & R. K. Koehn), pp. 165–190. Sunderland, MA: Sinauer Associates.
Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.
Nei, M. & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the USA* **76**, 5269–5273.
Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. & Bustamante, C. D. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research* **15**, 1566–1575.
Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics* **146**, 1501–1514.
Obbard, D. J., Welch, J. J., Kim, K. W. & Jiggins, F. M. (2009). Quantifying adaptive evolution in the Drosophila immune system. *Public Library of Science Genetics* **5**, e1000698.
Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**, 263–286.
Ometto, L., Glinka, S., De Lorenzo, D. & Stephan, W. (2005). Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Molecular Biology and Evolution* **22**, 2119–2130.
Orkin, S. H., Kazazian, H., Antonorakis, S. E., Goff, S. C., Boehm, C. D., Sextor, J. P., Waber, P. G., Giardina, P. J. V. (1982). Linkage of β-thalassemia mutations and β-globin gene polymorphisms with DNA polymorphisms in the human β-globin gene cluster. *Nature* **296**, 627–631.
Parmley, J. L. & Hurst, L. D. (2007). How do synonymous mutations affect fitness? *BioEssays* **29**, 515–519.
Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M. & Andolfatto, P. (2010). On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Molecular Biology and Evolution* **27**, 1226–1234.
Piganeau, G. & Eyre-Walker, A. (2003). Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *Proceedings of the National Academy of Sciences of the USA* **100**, 10335–10340.
Presgraves, D. (2005). Recombination enhances protein adaptation in *Drosophila melanogaster*. *Current Biology* **15**, 1651–1656.
Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69**, 124–137.
Provine, W. B. (1971). *The Origins of Theoretical Population Genetics*. Chicago, IL: University of Chicago Press.
Punnett, R. C. (1915). *Mimicry in Butterflies*. Cambridge: Cambridge University Press.
Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L. & Webster, M. T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. *Philosophical Transactions of the Royal Society B* **365**, 2571–2580.
Robertson, A. & Hill, W. G. (1983). Population and quantitative genetics of many linked loci in finite populations. *Proceedings of the Royal Society B* **219**, 253–263.
Rockman, M. V., Skrovanek, S. S. & Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* **330**, 372–376.
Sabeti, P. C., Varrilly, P., Fry, B., Lohnmuller, J., Hostetter, J., Cotsapas, C., Xie, X. H., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in the human genome. *Nature* **449**, 913–918.
Sackton, T. B., Kulathinal, R. J., Bergman, C. M., Quinlan, A. R., Dopman, E. B., Carneiro, M., Marth, G. T., Hartl, D. L., Clark, A. G. (2009). Population genomic inferences from sparse high-throughput sequencing of two populations of *Drosophila melanogaster*. *Genome Biology and Evolution* **1**, 449–450.
Sawyer, S. A., Parsch, J., Zhang, Z. & Hartl, D. L. (2007). Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proceedings of the National Academy of Sciences of the USA* **104**, 6504–6510.
Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. (2009). Comprehensive polymorphism study elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**, 342–345.
Schaeffer, S. W., Aquadro, C. F. & Langley, C. H. (1988). Restriction map variation in the *Notch* region of *Drosophila melanogaster*. *Molecular Biology and Evolution* **5**, 30–40.
Sella, G., Petrov, D. A., Przeworski, M. & Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? *Public Library of Science Genetics* **6**, e1000495.
Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H. Y., Hudson, R. R., Nielsen, R., Chen, Z., Wu, C. I. (2007). Adaptive genic evolution in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the USA* **104**, 2271–2276.
Sharp, P. M., Emery, L. B. & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B* **365**, 1203–1212.

Sharp, P. M. & Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **28**, 398–402.

Shiina, T., Ota, M., Shimizu, S., Katsuyama, Y., Hashimoto, N., Takasu, M., Gojobori, T., Inoko, H., Bahram S. (2006). Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics* **173**, 1555–1570.

Simmons, M. J. & Crow, J. F. (1977). Mutations affecting fitness in *Drosophila* populations. *Annual Review of Genetics* **11**, 49–78.

Simonsen, K. L., Churchill, G. A. & Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413–429.

Slatkin, M. & Wiehe, T. (1998). Genetic hitch-hiking in a subdivided population. *Genetical Research* **71**, 155–160.

Slotte, T., Foxe, J. P., Hazzouri, K. M. & Wright, S. I. (2010). Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. *Molecular Biology and Evolution* **27**, 1813–1821.

Small, K. S., Brudno, M., Hill, M. & Sidow, A. (2007). Extreme genomic variation in a natural population. *Proceedings of the National Academy of Sciences of the USA* **104**, 5698–5703.

Smith, N. G. C. & Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature* **415**, 1022–1024.

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B* **365**, 1245–1253.

Stephan, W. & Langley, C. H. (1989). Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. *Genetics* **121**, 89–99.

Sved, J. A. (1971). An estimate of heterosis in *Drosophila melanogaster*. *Genetical Research* **18**, 97–105.

The Encode Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 800–816.

Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y. D., Hudson, R. R. & Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the USA* **102**, 18508–18513.

Wakeley, J. 2008. *Coalescent Theory. An Introduction*. Greenwood Village, CO: Roberts & Co.

Welch, J. J. (2006). Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics* **173**, 821–827.

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D. & Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *Public Library of Science Genetics* **3**, 901–915.

Wright, A., Charlesworth, B., Rudan, I., Carothers, A. & Campbell, H. (2003). A polygenic basis for late-onset disease. *Trends in Genetics* **19**, 97–106.

Wright, S. (1922). *The Effects of Inbreeding and Cross-breeding on Guinea-pigs: I. Decline in Vigour*. *Bullletin 1090*. Washington, DC: U. S. Department of Agriculture.

Zeng, K. (2010). A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. *Molecular Biology and Evolution* **27**, 1327–1337.

Zeng, K. & Charlesworth, B. (2009). Estimating selection intensity on synonymous codon usage in a non-equilibrium population. *Genetics* **183**, 651–662.

Zeng, K. & Charlesworth, B. (2010). Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *Journal of Molecular Evolution* **70**, 116–128.

Zeng, K., Shi, S. & Wu, C. -I. (2007). Compound tests for the detection of hitchhiking under positive selection. *Molecular Biology and Evolution* **24**, 1898–1908.

Zhang, L. & Li, W.-H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Molecular Biology and Evolution* **22**, 2504–2507.